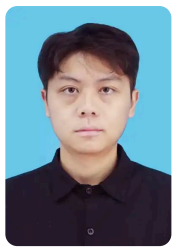
赵剑



Applying for: AI Application Engineer

13768710643 | 13768710643@163.com | sworddut | 广州

Education

中山大学 · 985 计算机技术 硕士 2024/09 - 2027/06

大连理工大学 · 985 网络工程 本科 2020/09 - 2024/06

获得奖项：第十四届蓝桥杯大赛国家级三等奖、大连理工大学“南国红豆奖学金”（连续三年）、学习优秀奖学金

Experience

上下文工程实习生 深信服科技 2026/01 - 2026/02

参与研发面向企业级销售场景的多 Agent 自动化系统（智能客服），针对长周期销售中话术非标、画像一致性差及评估困难等痛点，构建了从结构化思维链设计、动态记忆仲裁到自动化测评闭环的完整解决方案。

- **Multi-Agent 协作与结构化 CoT**：基于 **AutoGen** 设计任务调度与对话生成解耦的微服务架构；通过 **JSON 结构化思维链（Structured CoT）** 范式，强制模型遵循“意图识别-合规检查-策略生成”的推理路径，显著降低了销售话术生成的幻觉率，确保 SOP 节点流转逻辑可控、可解析。
- **动态画像记忆与冲突仲裁**：搭建基于 **Milvus** 的用户画像检索链路，针对多轮对话滑动窗口导致的画像信息冲突（如用户身份前后矛盾），设计基于 **Qwen-14B** 的**置信度打分仲裁机制**（Score > 8 触发更新），解决了 **90%** 以上的长文本画像震荡问题，保障了个性化推荐的精准度。
- **全链路监控与数据飞轮**：构建基于 **Apache Pulsar** 的非侵入式切面监控体系，实时采集业务指标；搭建“Badcase -> **Auto Prompt** -> Ground Truth 回归”的数据飞轮，实现了提示词工程的自动化迭代与分钟级效果评估。

大模型算法实习生 深圳清华大学先进材料与智能技术研究所 2025/04 - 2025/07

参与研发面向理科解题场景的垂直大语言模型，针对原始 PDF 教辅答案简略、推理过程缺失的问题，构建了从高并发数据处理、自动化清洗去重到 SFT+DPO 对齐训练的完整工程化流水线

- **高并发架构与开源贡献**：设计并实现基于 **Python Asyncio** 的高并发数据处理引擎，通过动态 Key 池管理将 DeepSeek/Qwen 等多模型请求吞吐提升至 **1000+ QPS**；将核心并发调度组件封装并**开源至 PyPI**。
- **智能清洗与向量去重**：搭建 **PDF-OCR-JSON** 自动化 ETL 链路，利用 Gemini 指令修复 OCR 噪声；针对 PDF 提取产生的冗余碎片，利用 **ChromaDB** 进行向量语义检索，基于题目相似度精准过滤约 **50%** 的重复内容，清洗出 **11万+** 条高质量 SFT 基座数据。
- **SFT+DPO 对齐**：针对理科解题中原始答案简略的问题，调用 Gemini-2.5-Pro 将原始答案重写为统一解题风格，并在 Unsloth 环境下执行LoRA SFT 与 DPO训练，最终解题准确率较基座提升**3%**，且稳定生成结构化解答过程。

前端开发实习生 深圳商汤科技有限公司（SenseTime） 2024/06 - 2024/09

- 参与公司核心 AI 平台 SenseCore 研发，基于 React + TypeScript 维护微前端子应用功能迭代。

Projects

MCP Edge Router：基于 Cloudflare Worker 的分布式工具调用网关 2025.05 - 至今

技术栈：Cloudflare Workers + FastMCP + Python + Node.js + JSON-RPC + OpenAI SDK

- 项目背景：在 AI Agent 场景下，工具规模增长会导致上下文膨胀与推理效率下降。设计并实现基于渐进式披露（Progressive Disclosure）的工具调用网关，按需暴露能力以提升多工具调用的稳定性与效率。
- 架构设计：采用 Agent → Edge Gateway → MCP Nodes 三层架构，在 Edge 层集中管理工具发现与路由，后端工具服务以独立 MCP 节点部署，实现能力解耦与横向扩展。
- 按需工具发现与精确路由：Agent 仅在需要时发现目标节点及其工具，避免一次性暴露完整工具空间，有效控制上下文规模并降低误调用风险。

Research

Towards Culturally Fair Multimodal Generation (ACM Multimedia 2025) Yifan Zeng, Fangzhou Dong, *Jian Zhao*

- 参与文生视频（T2V）模型文化偏见的量化评估实验，负责部分实验代码实现与自动化评测流程，支持多模型、多文化维度的对比分析。

Skills

- **LLM Engineering:** LLM Fine-tuning (SFT/DPO), RAG, Agent, PyTorch, Unsloth, LangChain, ChromaDB.
- **Backend & System**: Python, Node.js, Linux, Git, MongoDB, Redis, Docker.
- **Frontend:** Vue3, React, TypeScript